Data Profiling and Data Quality Checks in SAS

Using SAS Studio on SAS On Demand for Academics (SODA)

Imported Your Data Already?

- If you already have your data in SAS Studio on SAS On Demand for Academics, you can skip the slides providing an overview of the import process.

Dataset

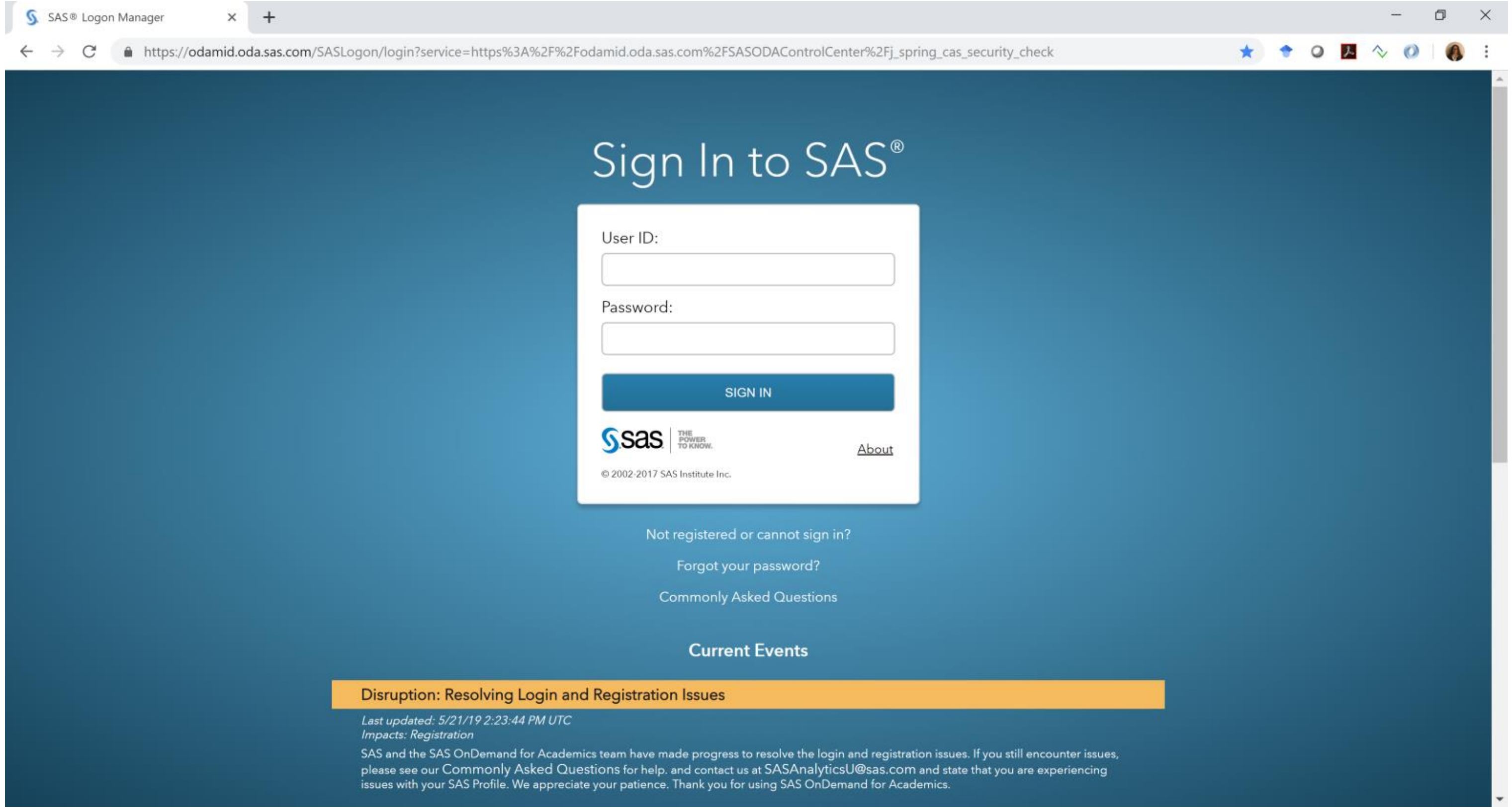
- This tutorial is a walkthrough with a sample set of data. You may use this to walk through the tutorial, if you wish, but for your assignments, you will be asked to use your own dataset (as specified within the course).

Dataset reference:

Skoryk, M. (2021). Sepsis Prediction from Clinical Data. Version 1. Retrieved from https://www.kaggle.com/maxskoryk/datasepsis

Access the SAS OnDemand for Academics Control Center

https://odamid.oda.sas.com/SASODAControlCenter


SAS® Logon Manager
https://odamid.oda.sas.com/SASLogon/login?service=https%3A%2F%2Fodamid.oda.sas.com%2FSASODAControlCenter%2Fj_spring_cas_security_check

Sign In to SAS®

User ID:

Password:

SIGN IN

About

© 2002-2017 SAS Institute Inc.

Not registered or cannot sign in?

Forgot your password?

Commonly Asked Questions

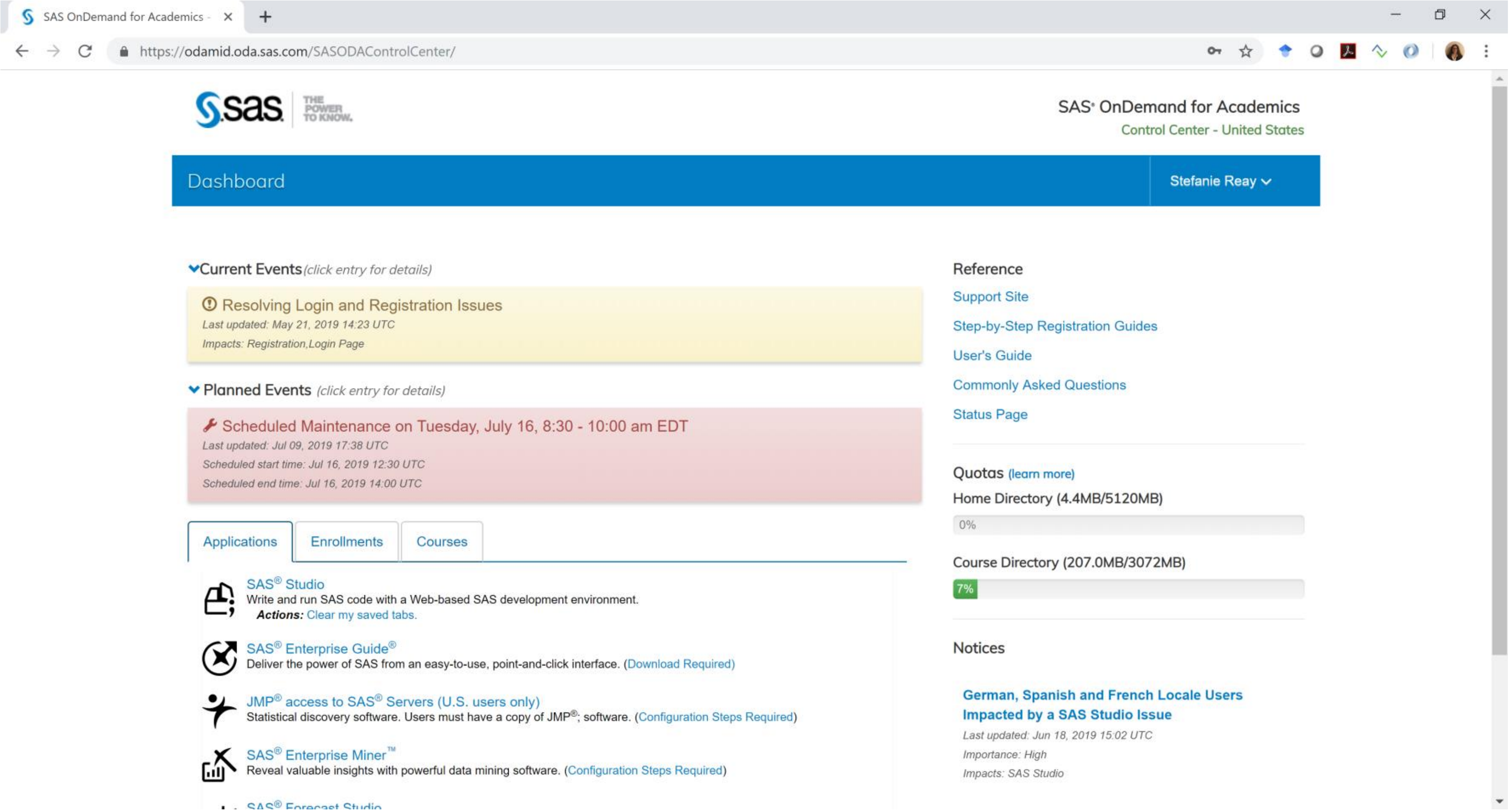
Current Events

Disruption: Resolving Login and Registration Issues

Last updated: 5/21/19 2:23:44 PM UTC
Impacts: Registration

SAS and the SAS OnDemand for Academics team have made progress to resolve the login and registration issues. If you still encounter issues, please see our Commonly Asked Questions for help. and contact us at SASAnalyticsU@sas.com and state that you are experiencing issues with your SAS Profile. We appreciate your patience. Thank you for using SAS OnDemand for Academics.

SAS OnDemand for Academics (SODA) Control Center


SAS OnDemand for Academics -
https://odamid.oda.sas.com/SASODAControlCenter/
sas THE POWER TO KNOW.
Dashboard
Stefanie Reay
Applications
Enrollments
Courses

SAS® OnDemand for Academics
Control Center - United States

Current Events *(click entry for details)*

Resolving Login and Registration Issues
Last updated: May 21, 2019 14:23 UTC
Impacts: Registration,Login Page

Planned Events *(click entry for details)*

Scheduled Maintenance on Tuesday, July 16, 8:30 - 10:00 am EDT
Last updated: Jul 09, 2019 17:38 UTC
Scheduled start time: Jul 16, 2019 12:30 UTC
Scheduled end time: Jul 16, 2019 14:00 UTC

SAS® Studio
Write and run SAS code with a Web-based SAS development environment.
Actions: Clear my saved tabs.

SAS® Enterprise Guide®
Deliver the power of SAS from an easy-to-use, point-and-click interface. (Download Required)

JMP® access to SAS® Servers (U.S. users only)
Statistical discovery software. Users must have a copy of JMP®; software. (Configuration Steps Required)

SAS® Enterprise Miner™
Reveal valuable insights with powerful data mining software. (Configuration Steps Required)

SAS® Forecast Studio

Reference

- Support Site
- Step-by-Step Registration Guides
- User's Guide
- Commonly Asked Questions
- Status Page

Quotas (learn more)

Home Directory (4.4MB/5120MB)
0%

Course Directory (207.0MB/3072MB)
7%

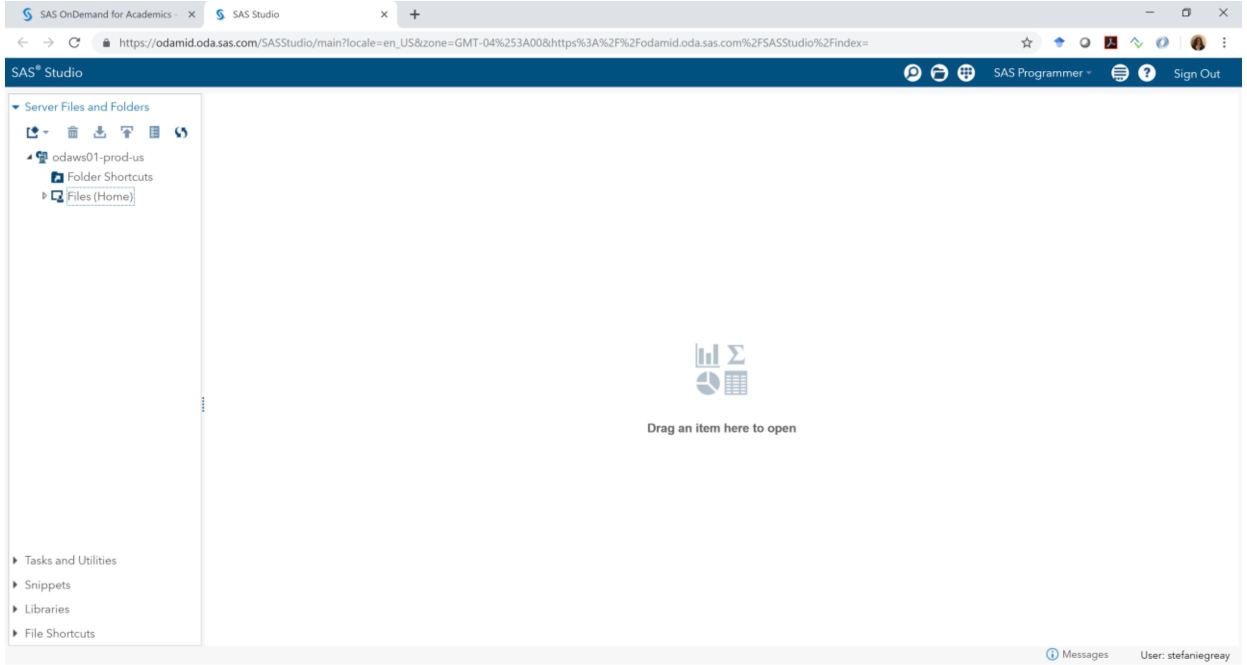
Notices

German, Spanish and French Locale Users Impacted by a SAS Studio Issue
Last updated: Jun 18, 2019 15:02 UTC
Importance: High
Impacts: SAS Studio

© Stefanie G. Reay, MS, PhD, Capella University, 2020

SAS Studio

https://odamid.oda.sas.com/SASStudio/

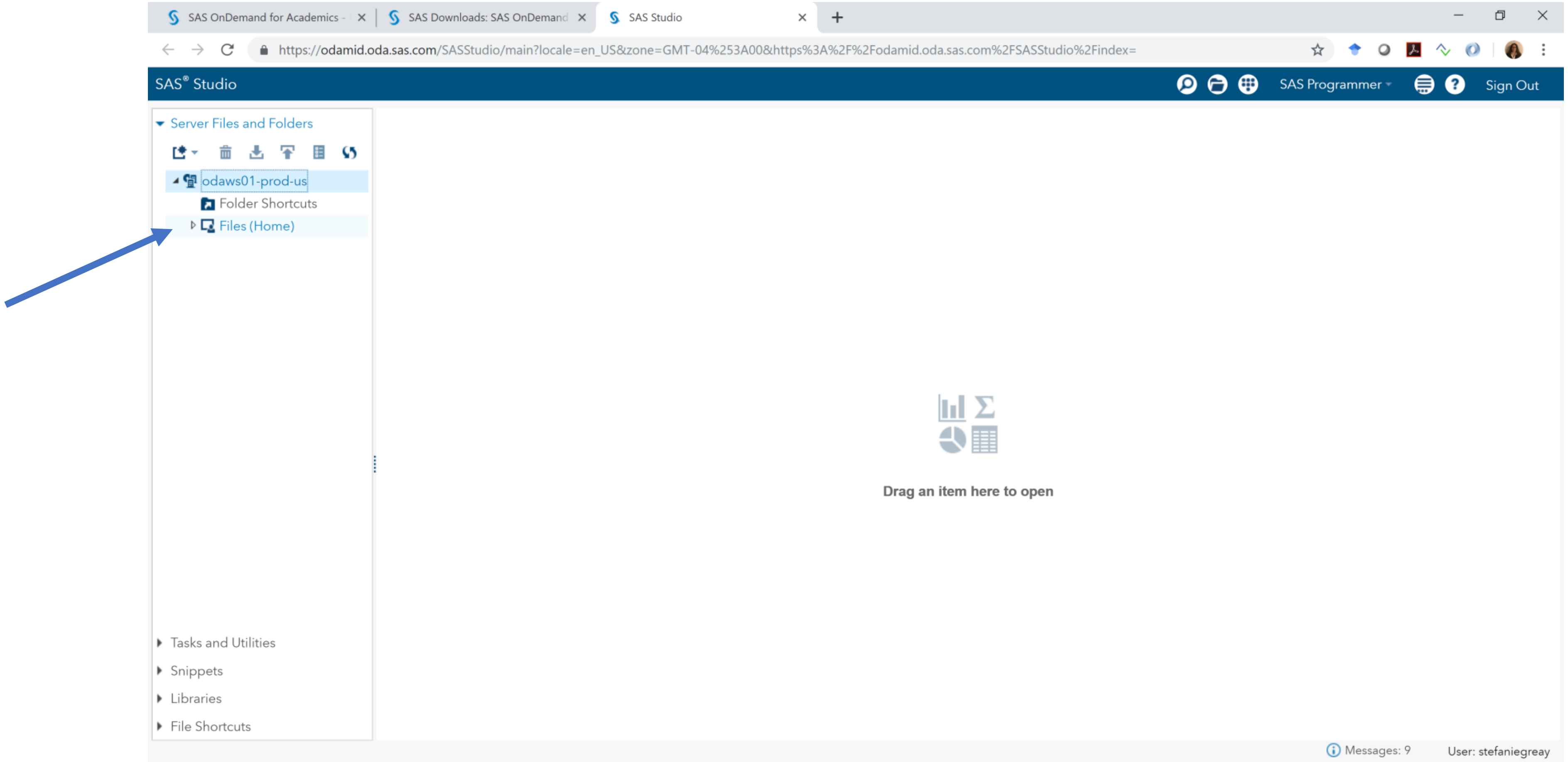


https://odamid.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-04%253A00&https%3A%2F%2Fodamid.oda.sas.com%2FSASStudio%2Findex=

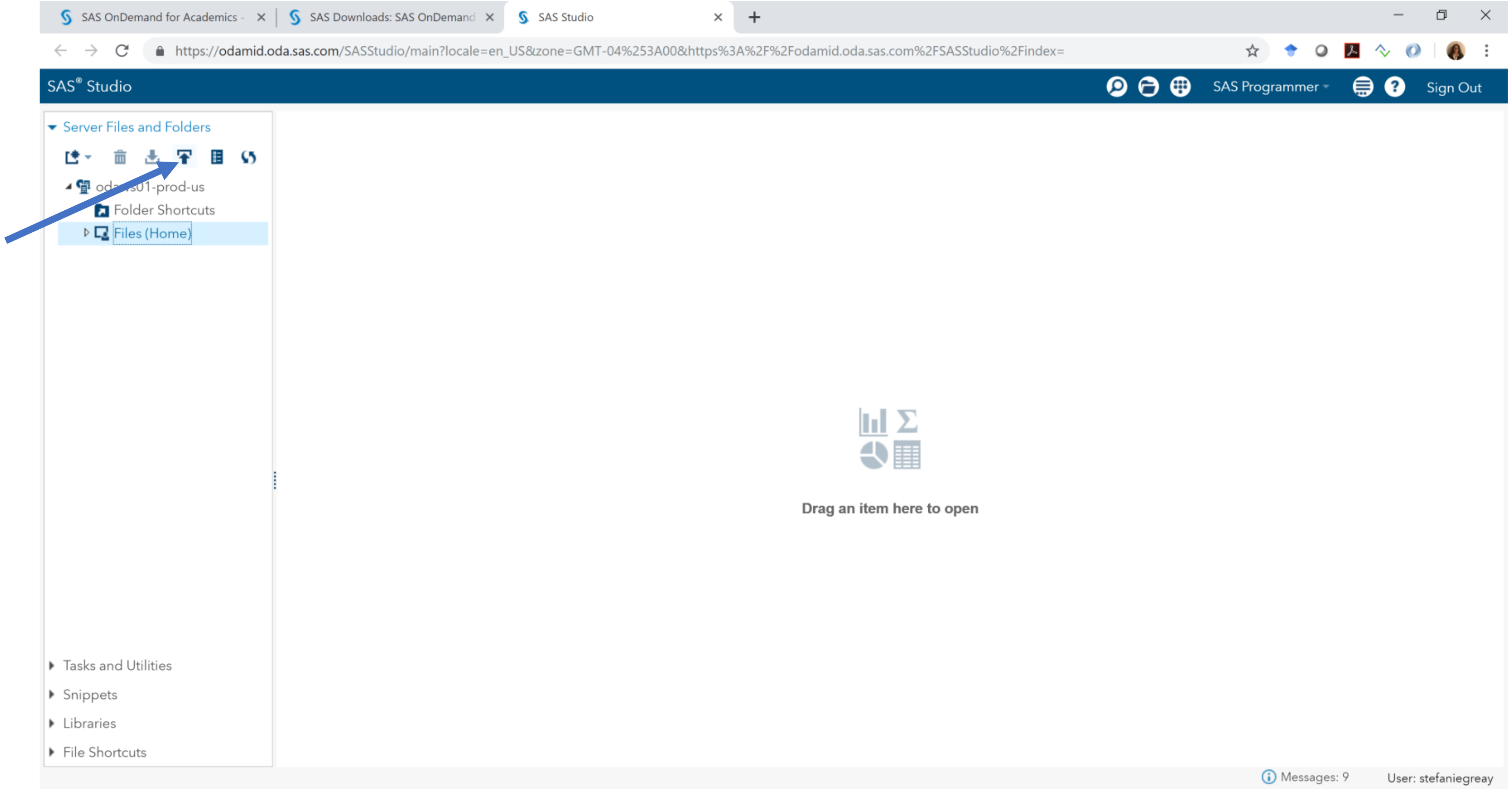
Drag an item here to open

User: stefaniegreay

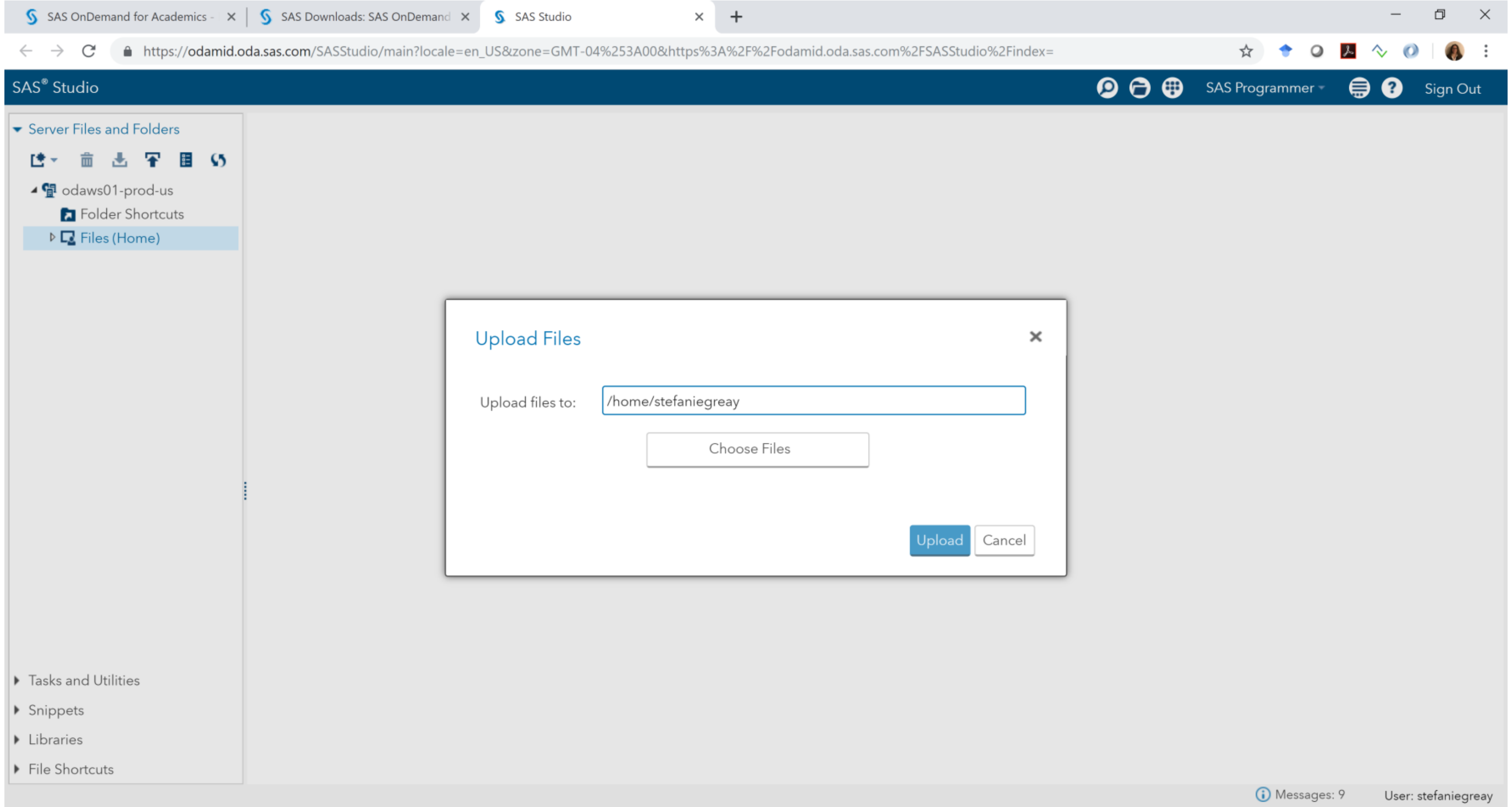
Click on Files(Home)


SAS OnDemand for Academics
SAS Downloads: SAS OnDemand
SAS Studio
https://odamid.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-04%253A00&https%3A%2F%2Fodamid.oda.sas.com%2FSASStudio%2Findex=
SAS® Studio
SAS Programmer
Sign Out
Server Files and Folders
odaws01-prod-us
Folder Shortcuts
Files (Home)
Drag an item here to open
Tasks and Utilities
Snippets
Libraries
File Shortcuts
Messages: 9
User: stefaniegreay

The Upload button will display in dark blue


SAS OnDemand for Academics
SAS Downloads: SAS OnDemand
SAS Studio
https://odamid.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT-04%253A00&https%3A%2F%2Fodamid.oda.sas.com%2FSASStudio%2Findex=
SAS® Studio
SAS Programmer
Sign Out
Server Files and Folders
odaws01-prod-us
Folder Shortcuts
Files (Home)
Drag an item here to open
Tasks and Utilities
Snippets
Libraries
File Shortcuts
Messages: 9
User: stefaniegreay

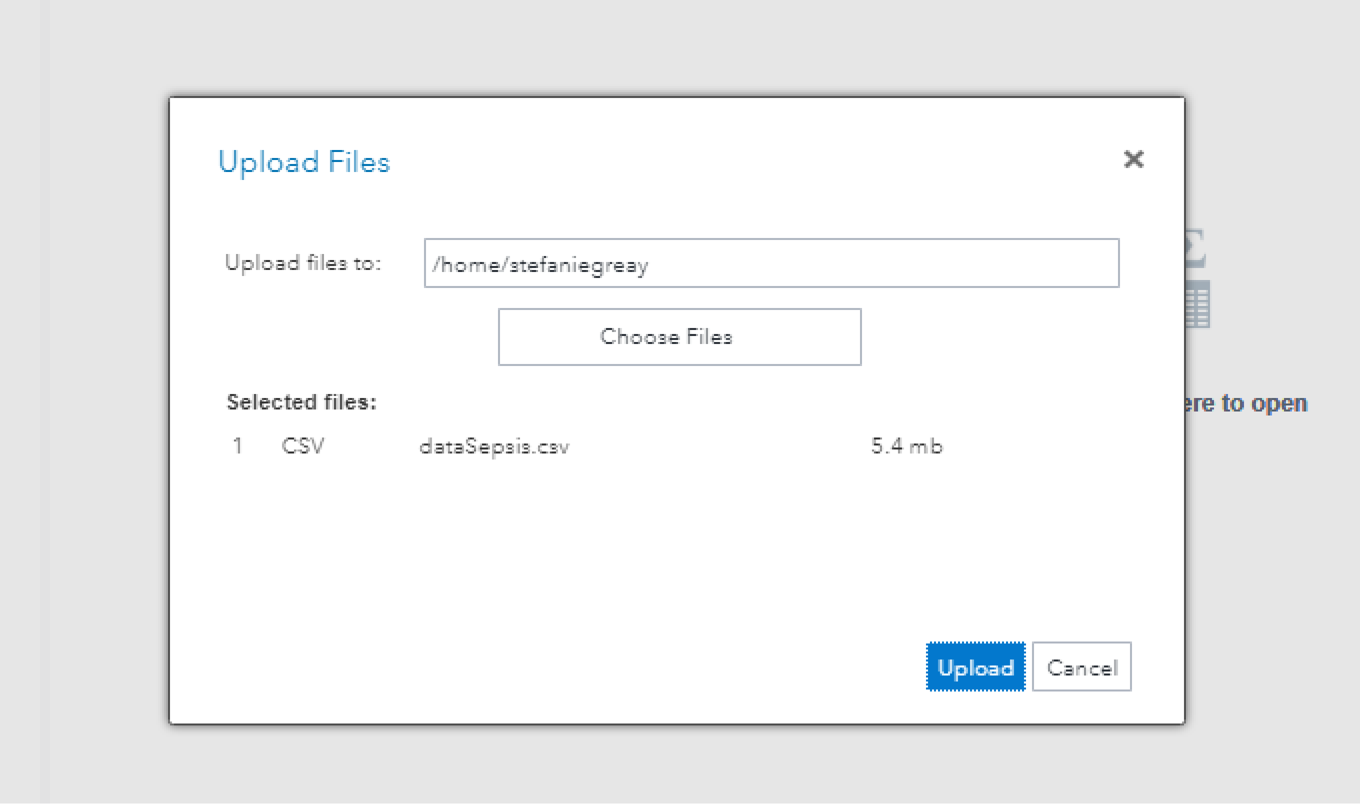
You can create a folder at this point, if you wish, or simply upload to your home directory.



Upload Files

Upload files to: /home/stefaniegreay

Select “Choose Files” to browse your computer for the dataset you want to upload. Once the dataset has been selected, click “Upload.”



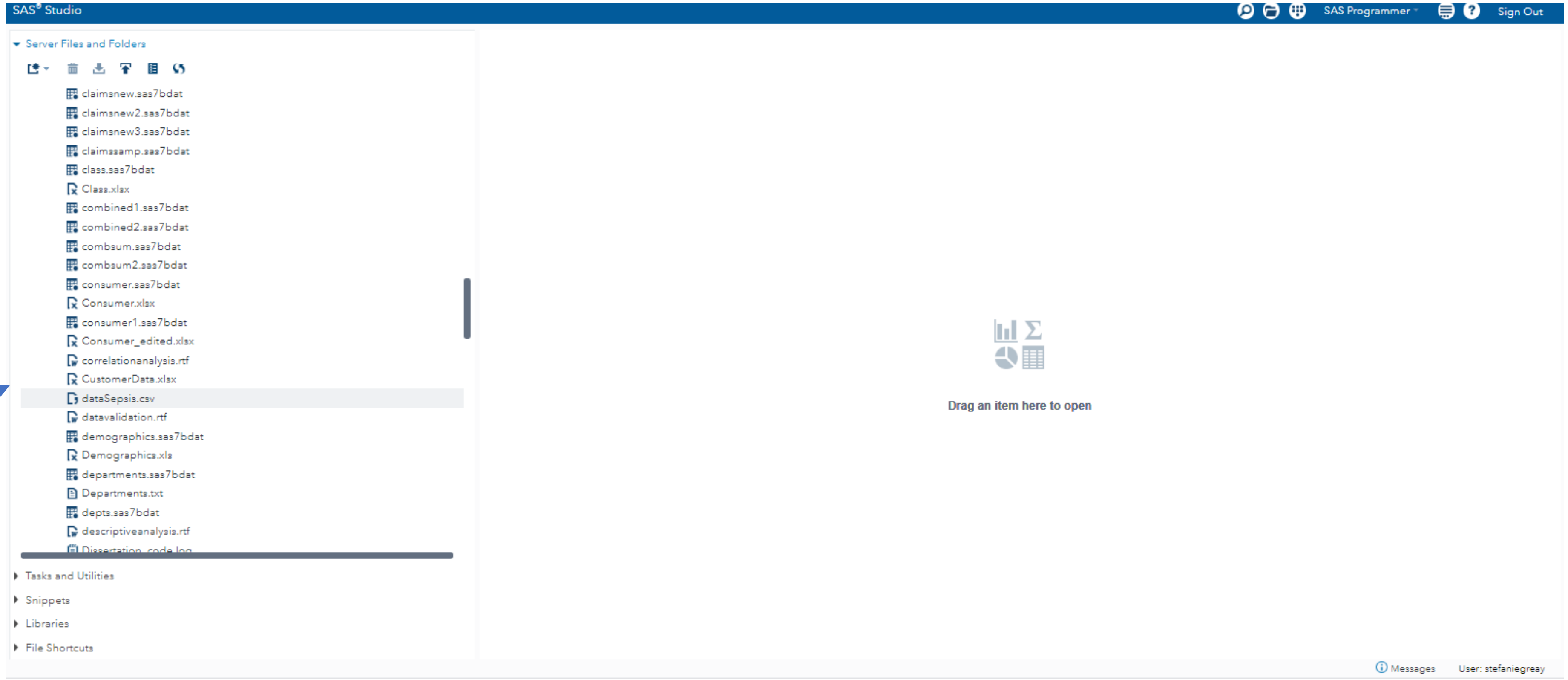
Upload Files

Upload files to: /home/stefaniegreay

Selected files:

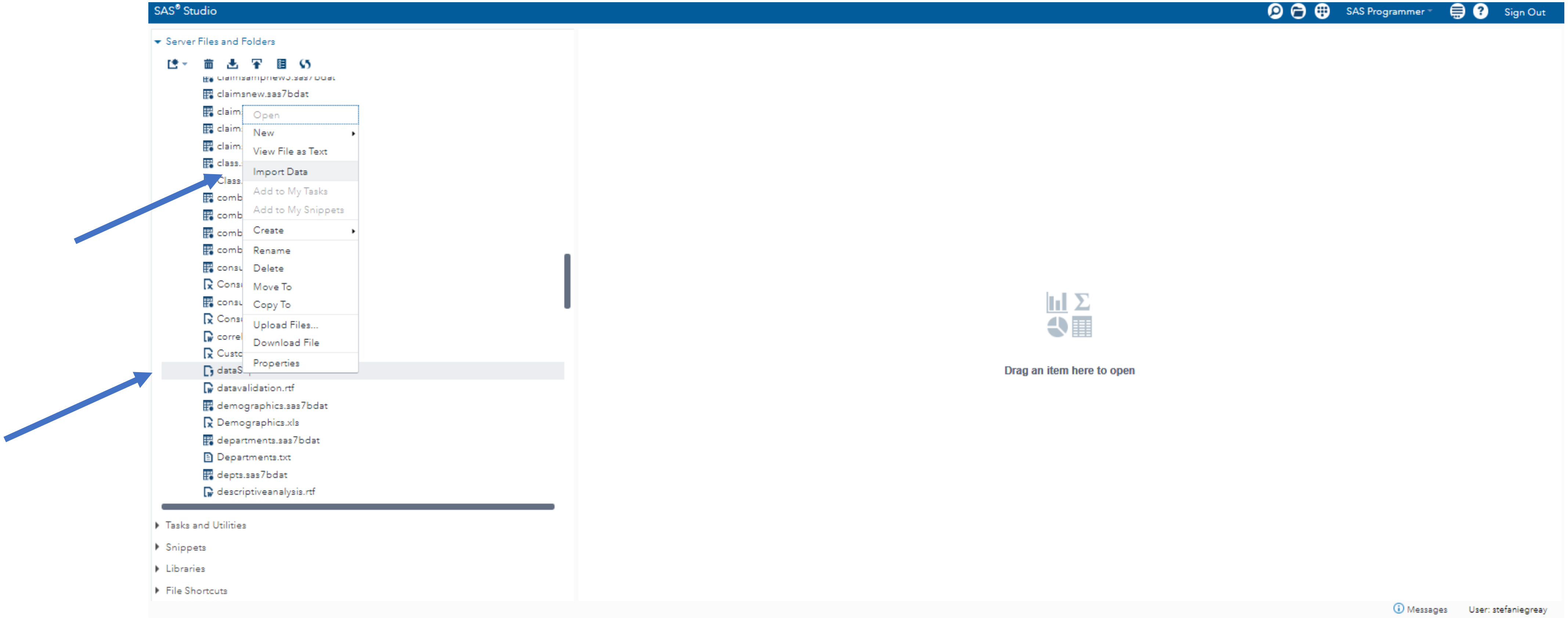
1 CSV dataSepsis.csv 5.4 mb

You will be able to view your files by clicking on “Files(Home)” to verify that your file successfully uploaded.



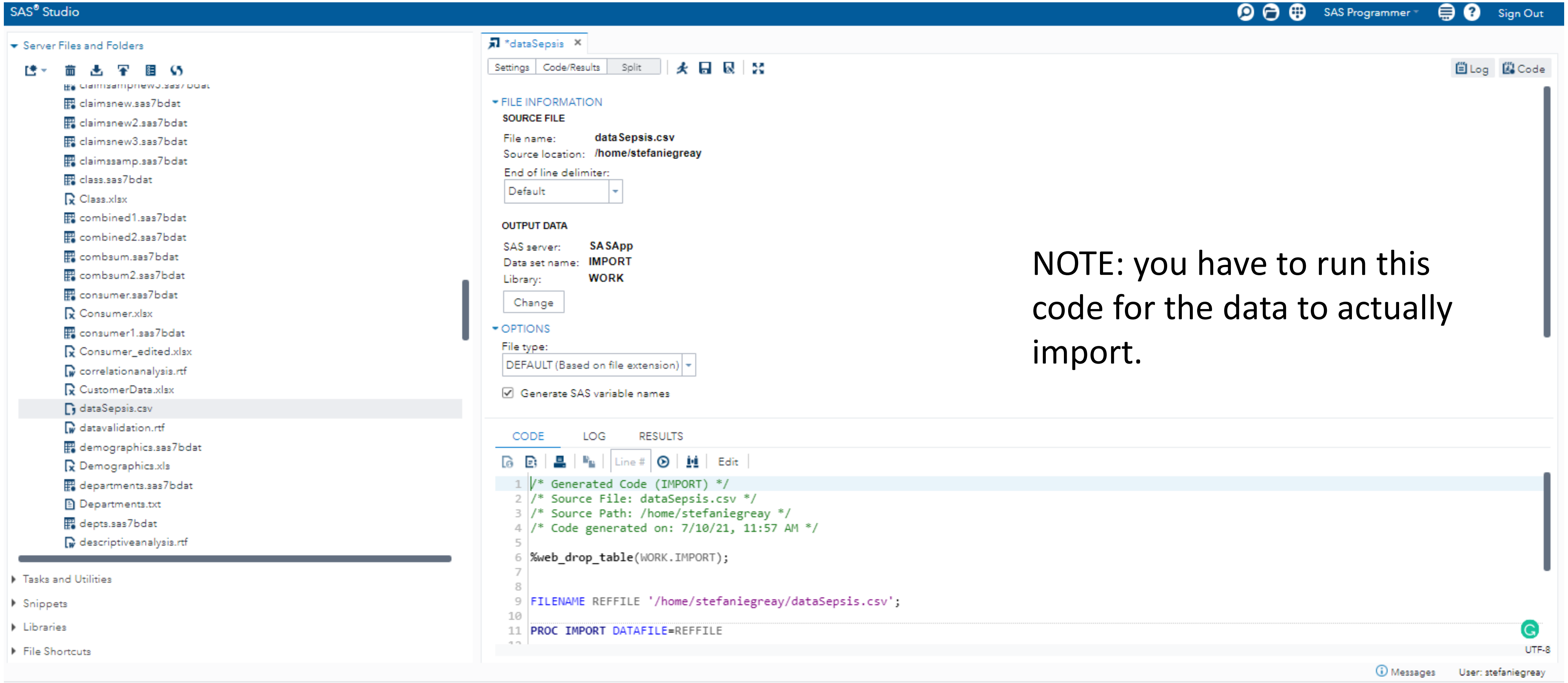
claimsnew.sas7bdat
claimsnew2.sas7bdat
claimsnew3.sas7bdat
claimssamp.sas7bdat
class.sas7bdat
Class.xlsx
combined1.sas7bdat
combined2.sas7bdat
combsum.sas7bdat
combsum2.sas7bdat
consumer.sas7bdat
Consumer.xlsx
consumer1.sas7bdat
Consumer_edited.xlsx
correlationanalysis.rtf
CustomerData.xlsx
dataSepsis.csv
datavalidation.rtf
demographics.sas7bdat
Demographics.xls
departments.sas7bdat
Departments.txt
depts.sas7bdat
descriptiveanalysis.rtf
Dissertation_code.log

To import the dataset into a SAS dataset format (from the current csv format), right click on the name of the file, and select “Import Data.”



© Stefanie G. Reay, MS, PhD, Capella University, 2020

The Proc Import code will be written for you (save this as a template to use for future imports!)



FILE INFORMATION

SOURCE FILE

File name: dataSepsis.csv

Source location: /home/stefaniegreay

End of line delimiter: Default

OUTPUT DATA

SAS server: SASApp

Data set name: IMPORT

Library: WORK

OPTIONS

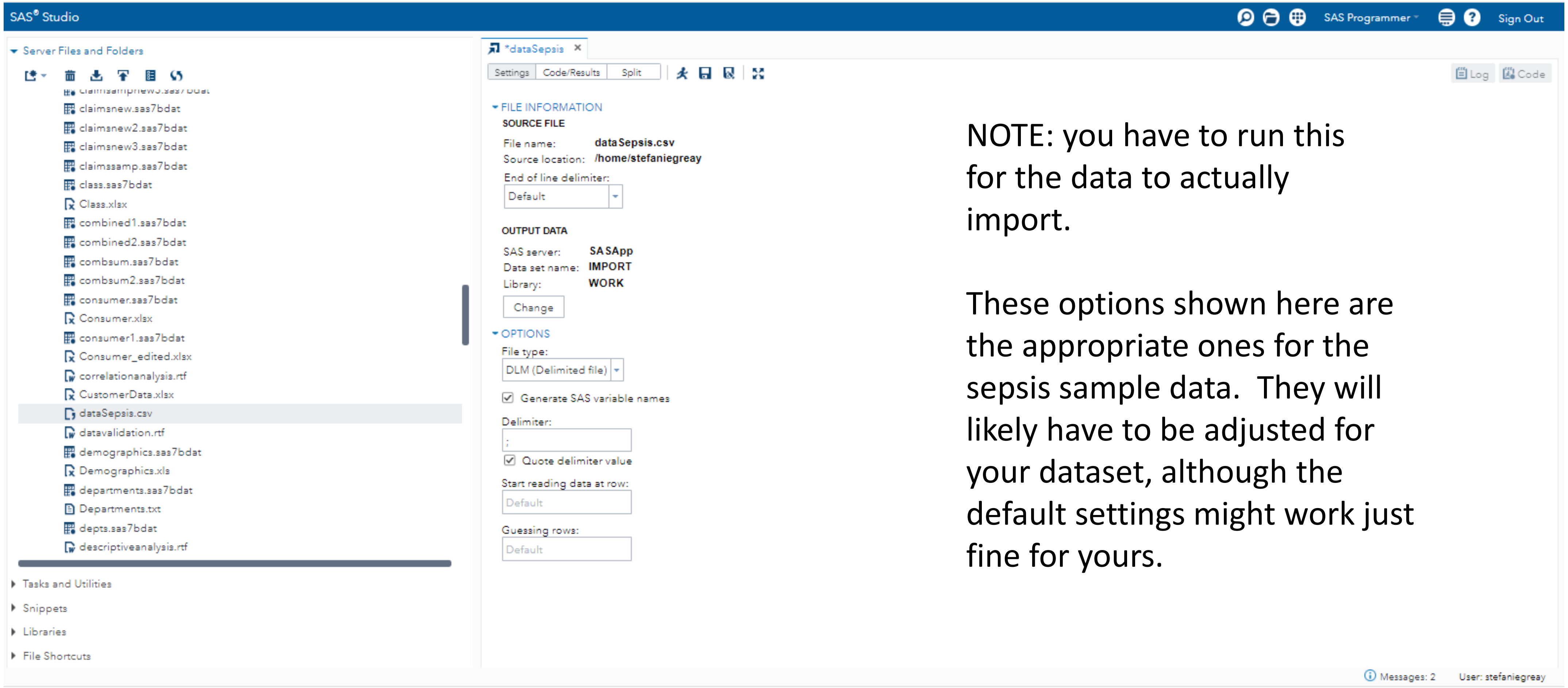
File type: DEFAULT (Based on file extension)

☑ Generate SAS variable names

```
/* Generated Code (IMPORT) */
/* Source File: dataSepsis.csv */
/* Source Path: /home/stefaniegreay */
/* Code generated on: 7/10/21, 11:57 AM */

%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/home/stefaniegreay/dataSepsis.csv';

PROC IMPORT DATAFILE=REFFILE
```

NOTE: you have to run this code for the data to actually import.

The Proc Import code will be written for you (save this as a template to use for future imports!)



FILE INFORMATION

SOURCE FILE

File name: dataSepsis.csv

Source location: /home/stefaniegreay

End of line delimiter: Default

OUTPUT DATA

SAS server: SASApp

Data set name: IMPORT

Library: WORK

OPTIONS

File type: DLM (Delimited file)

☑ Generate SAS variable names

Delimiter: ;

☑ Quote delimiter value

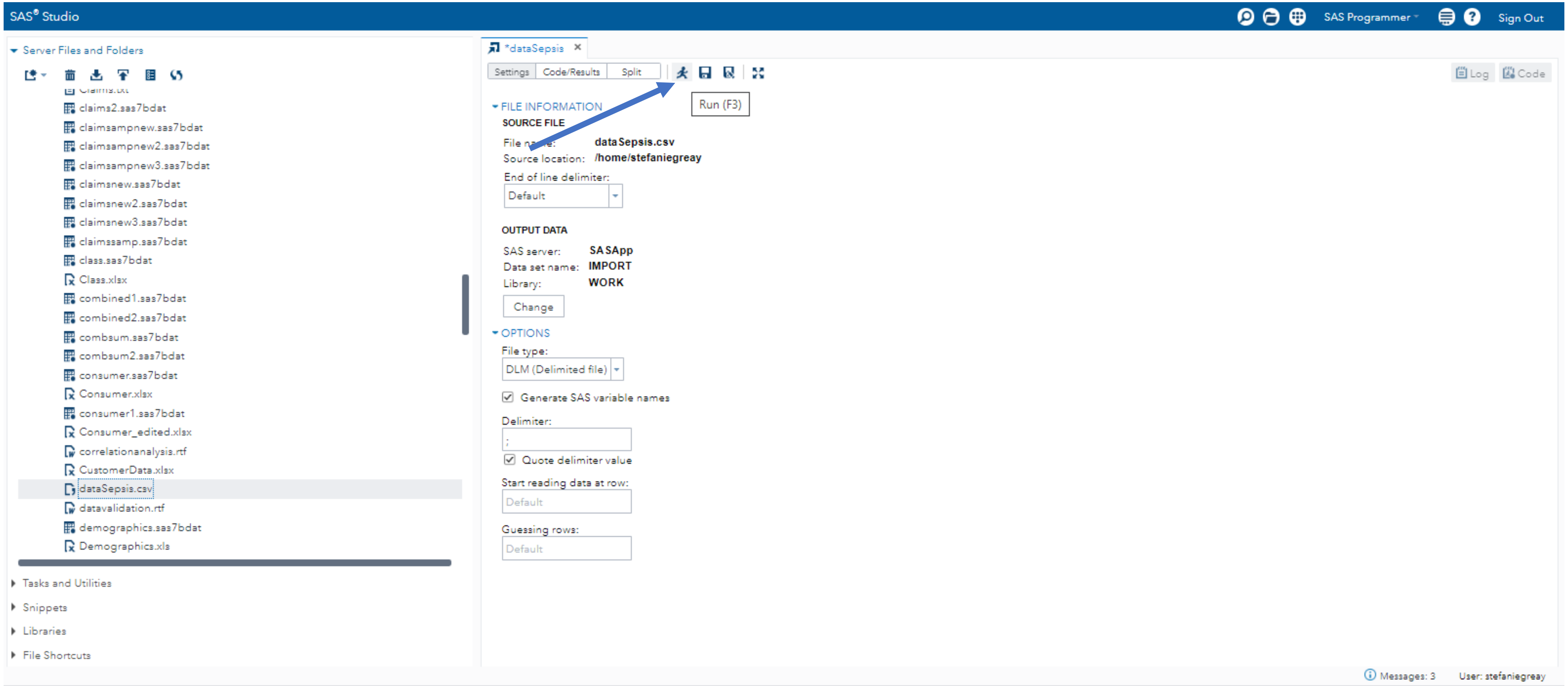
Start reading data at row: Default

Guessing rows: Default

NOTE: you have to run this for the data to actually import.

These options shown here are the appropriate ones for the sepsis sample data. They will likely have to be adjusted for your dataset, although the default settings might work just fine for yours.

To run the code, click the icon that looks like a guy running.



FILE INFORMATION

SOURCE FILE

File name: dataSepsis.csv

Source location: /home/stefaniegreay

End of line delimiter: Default

OUTPUT DATA

SAS server: SASApp

Data set name: IMPORT

Library: WORK

Change

OPTIONS

File type: DLM (Delimited file)

☑ Generate SAS variable names

Delimiter: ;

☑ Quote delimiter value

Start reading data at row: Default

Guessing rows: Default

Messages: 3 User: stefaniegreay

When you run the import, you will see the dataset and summary in the output data window when you click the “Code/Results” or “Split” tab and then “Output Data” and can verify its success.

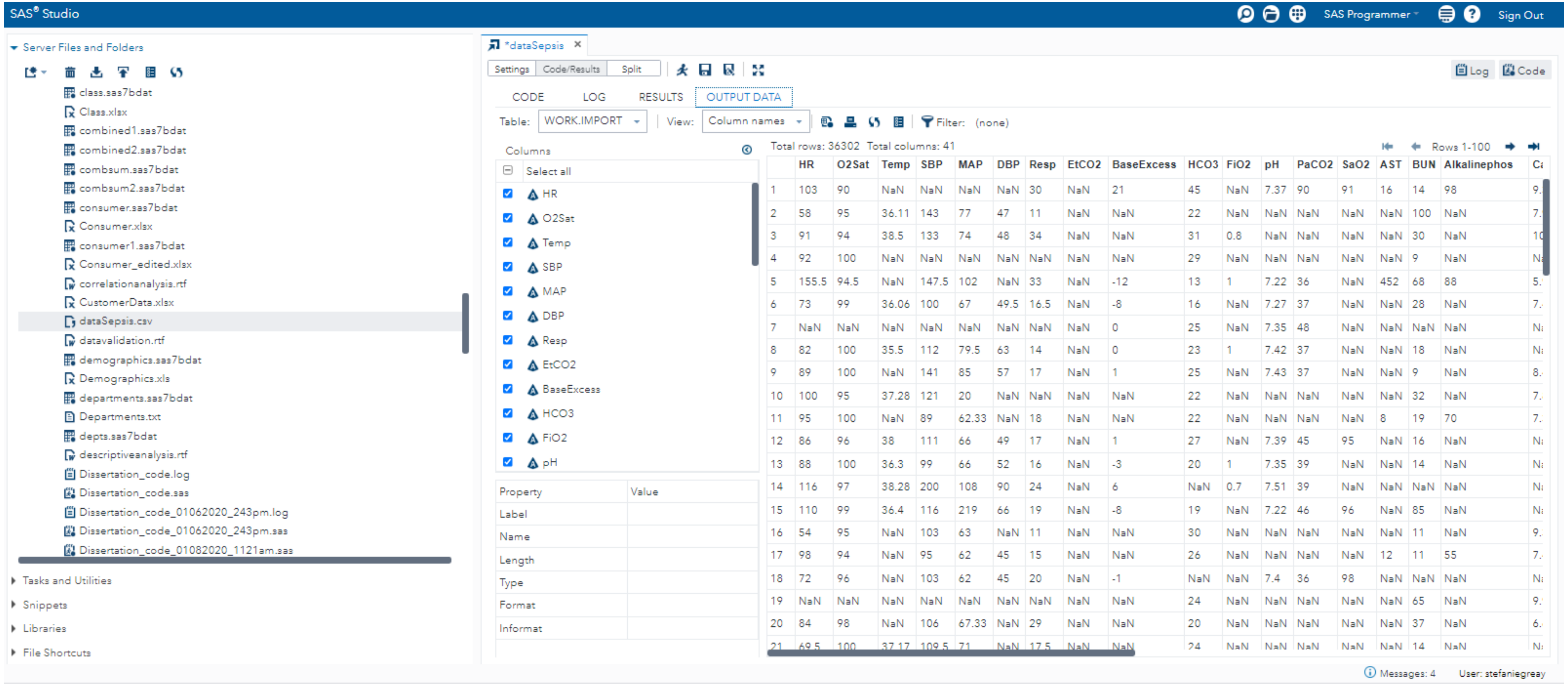


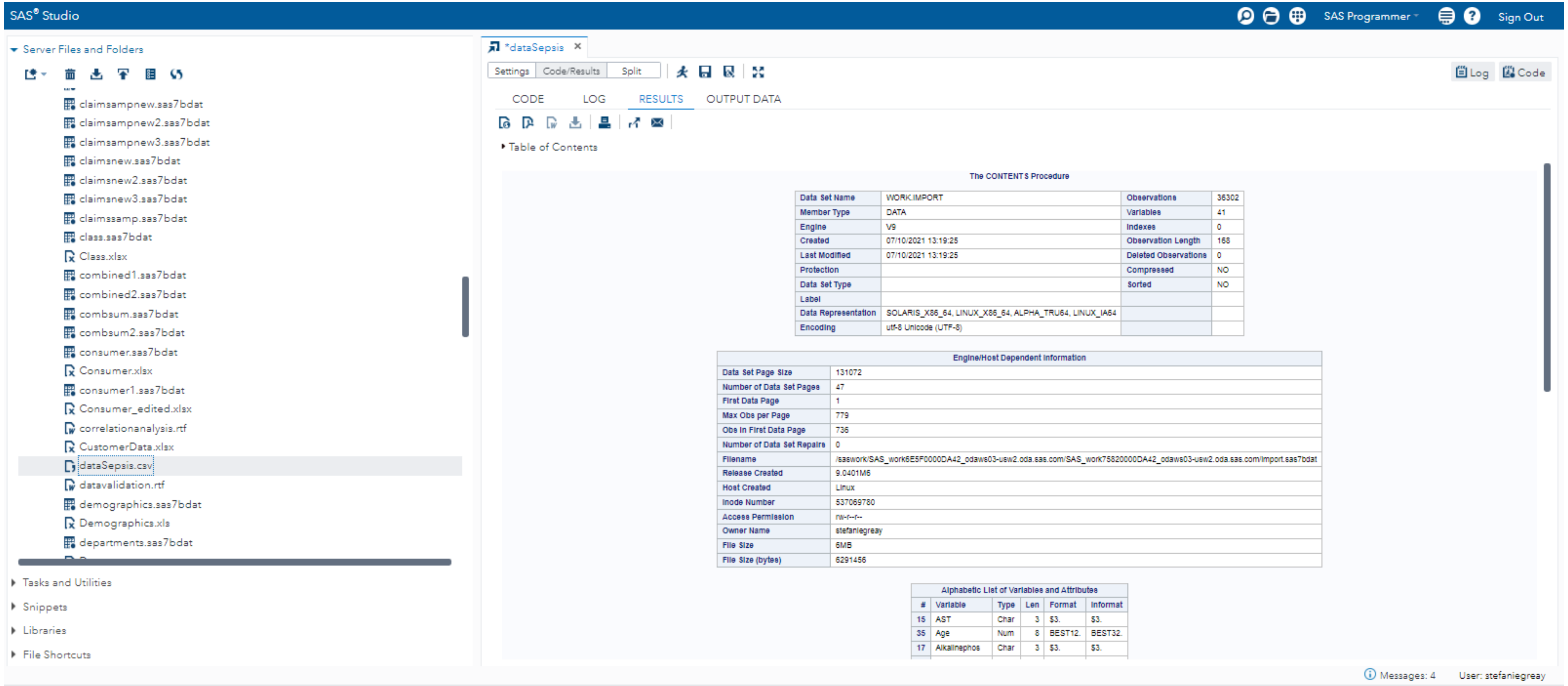
Table: WORK.IMPORT View: Column names Filter: (none)

Total rows: 36302 Total columns: 41

	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	FiO2	pH	PaCO2	SaO2	AST	BUN	Alkalinephos	Ca
1	103	90	NaN	NaN	NaN	NaN	30	NaN	21	45	NaN	7.37	90	91	16	14	98	9.
2	58	95	36.11	143	77	47	11	NaN	NaN	22	NaN	NaN	NaN	NaN	NaN	100	NaN	7.
3	91	94	38.5	133	74	48	34	NaN	NaN	31	0.8	NaN	NaN	NaN	NaN	30	NaN	10
4	92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	29	NaN	NaN	NaN	NaN	NaN	9	NaN	N
5	155.5	94.5	NaN	147.5	102	NaN	33	NaN	-12	13	1	7.22	36	NaN	452	68	88	5.
6	73	99	36.06	100	67	49.5	16.5	NaN	-8	16	NaN	7.27	37	NaN	NaN	28	NaN	7.
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0	25	NaN	7.35	48	NaN	NaN	NaN	NaN	N
8	82	100	35.5	112	79.5	63	14	NaN	0	23	1	7.42	37	NaN	NaN	18	NaN	N
9	89	100	NaN	141	85	57	17	NaN	1	25	NaN	7.43	37	NaN	NaN	9	NaN	8.
10	100	95	37.28	121	20	NaN	NaN	NaN	NaN	22	NaN	NaN	NaN	NaN	NaN	32	NaN	7.
11	95	100	NaN	89	62.33	NaN	18	NaN	NaN	22	NaN	NaN	NaN	NaN	8	19	70	7.
12	86	96	38	111	66	49	17	NaN	1	27	NaN	7.39	45	95	NaN	16	NaN	N
13	88	100	36.3	99	66	52	16	NaN	-3	20	1	7.35	39	NaN	NaN	14	NaN	N
14	116	97	38.28	200	108	90	24	NaN	6	NaN	0.7	7.51	39	NaN	NaN	NaN	NaN	N
15	110	99	36.4	116	219	66	19	NaN	-8	19	NaN	7.22	46	96	NaN	85	NaN	N
16	54	95	NaN	103	63	NaN	11	NaN	NaN	30	NaN	NaN	NaN	NaN	NaN	11	NaN	9.
17	98	94	NaN	95	62	45	15	NaN	NaN	26	NaN	NaN	NaN	NaN	12	11	55	7.
18	72	96	NaN	103	62	45	20	NaN	-1	NaN	NaN	7.4	36	98	NaN	NaN	NaN	N
19	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	24	NaN	NaN	NaN	NaN	NaN	65	NaN	9.
20	84	98	NaN	106	67.33	NaN	29	NaN	NaN	20	NaN	NaN	NaN	NaN	NaN	37	NaN	6.
21	69.5	100	37.17	109.5	71	NaN	17.5	NaN	NaN	24	NaN	NaN	NaN	NaN	NaN	14	NaN	N

Messages: 4 User: stefaniegreay

When you click the “Code/Results” or “Split” tab and then “Results,” you can see the contents of the dataset, to verify the number of observations and variables are as expected.



The CONTENTS Procedure

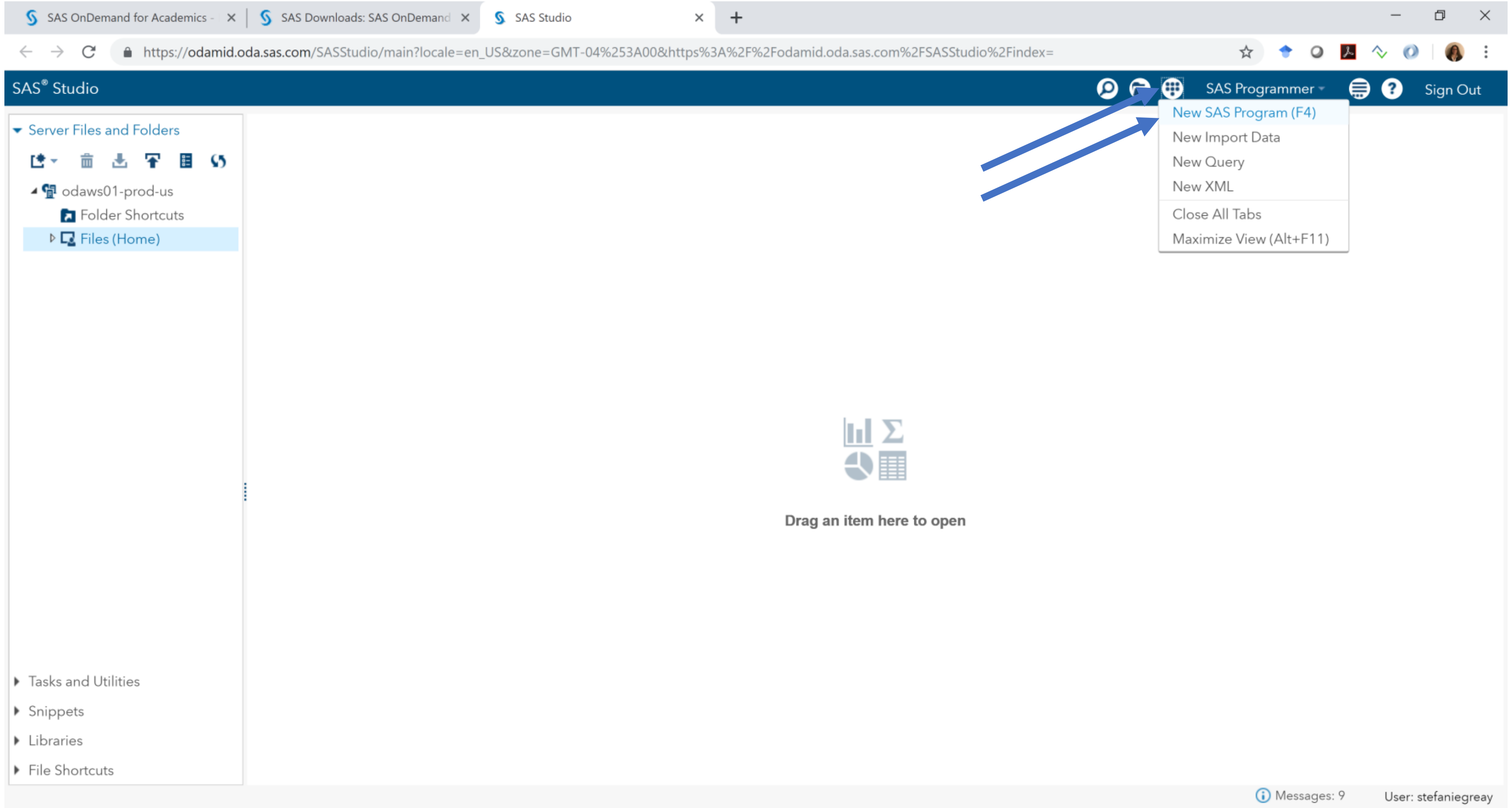
Data Set Name	WORK.IMPORT	Observations	36302
Member Type	DATA	Variables	41
Engine	V9	Indexes	0
Created	07/10/2021 13:19:25	Observation Length	168
Last Modified	07/10/2021 13:19:25	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	47
First Data Page	1
Max Obs per Page	779
Obs in First Data Page	736
Number of Data Set Repairs	0
Filename	/saswork/SAS_work6E5F0000DA42_odaws03-usw2.oda.sas.com/SAS_work75820000DA42_odaws03-usw2.oda.sas.com/import.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	537069780
Access Permission	rw-r--r--
Owner Name	stefaniegreay
File Size	6MB
File Size (bytes)	6291456

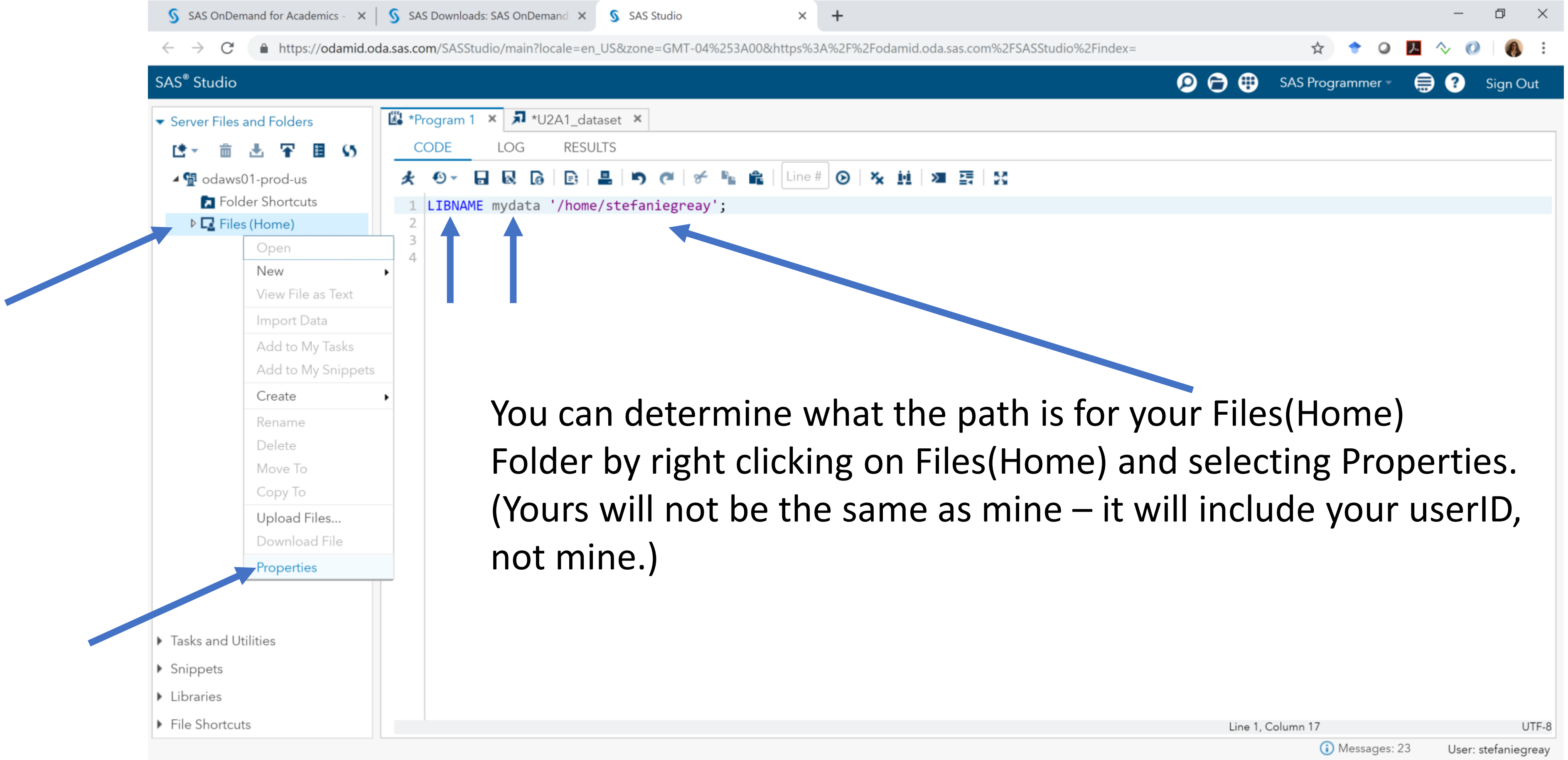
Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
15	AST	Char	3	$3.	$3.
35	Age	Num	8	BEST12.	BEST32.
17	Alkalinephos	Char	3	$3.	$3.

To get started working with the dataset you just imported, start a new SAS program.



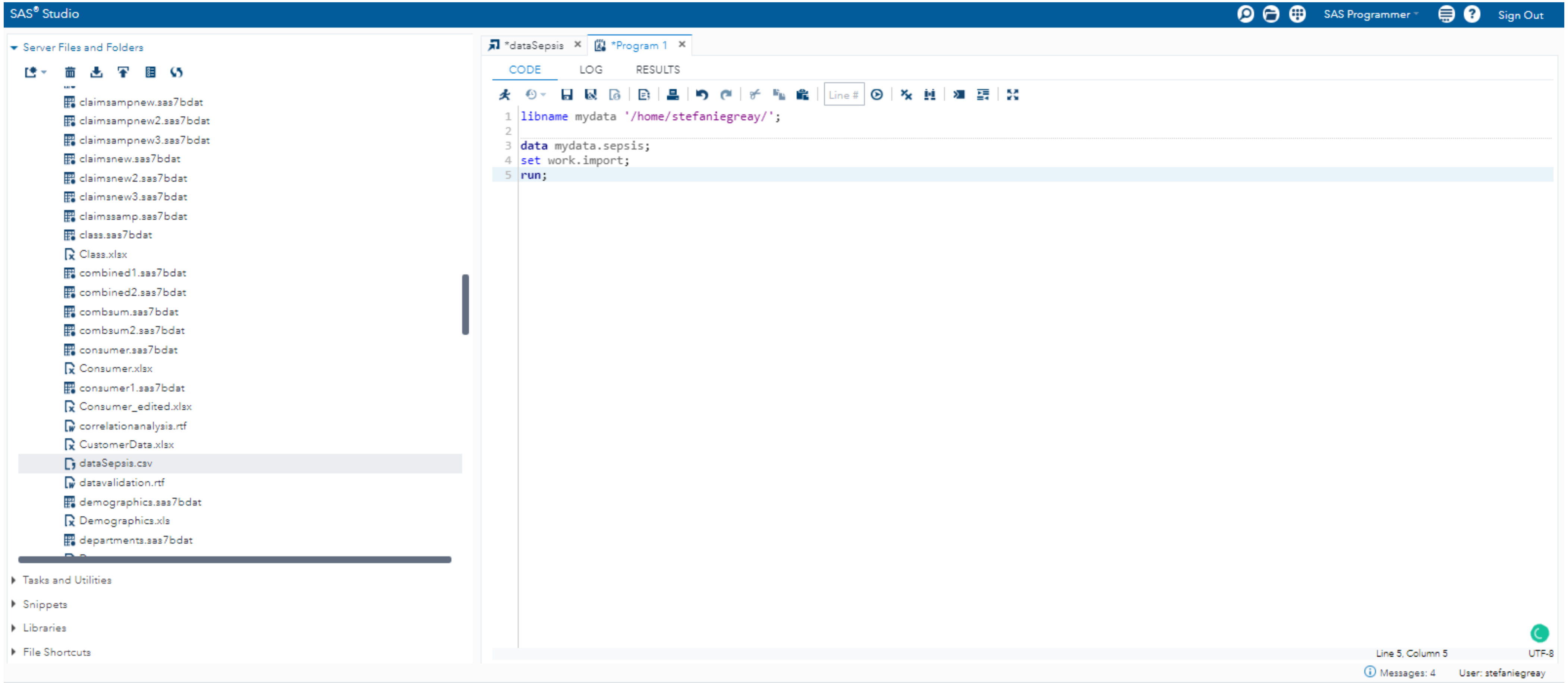
To create a SAS Library for your Files(Home) folder, you need to use a libname statement



```
LIBNAME mydata '/home/stefaniegreay';
```

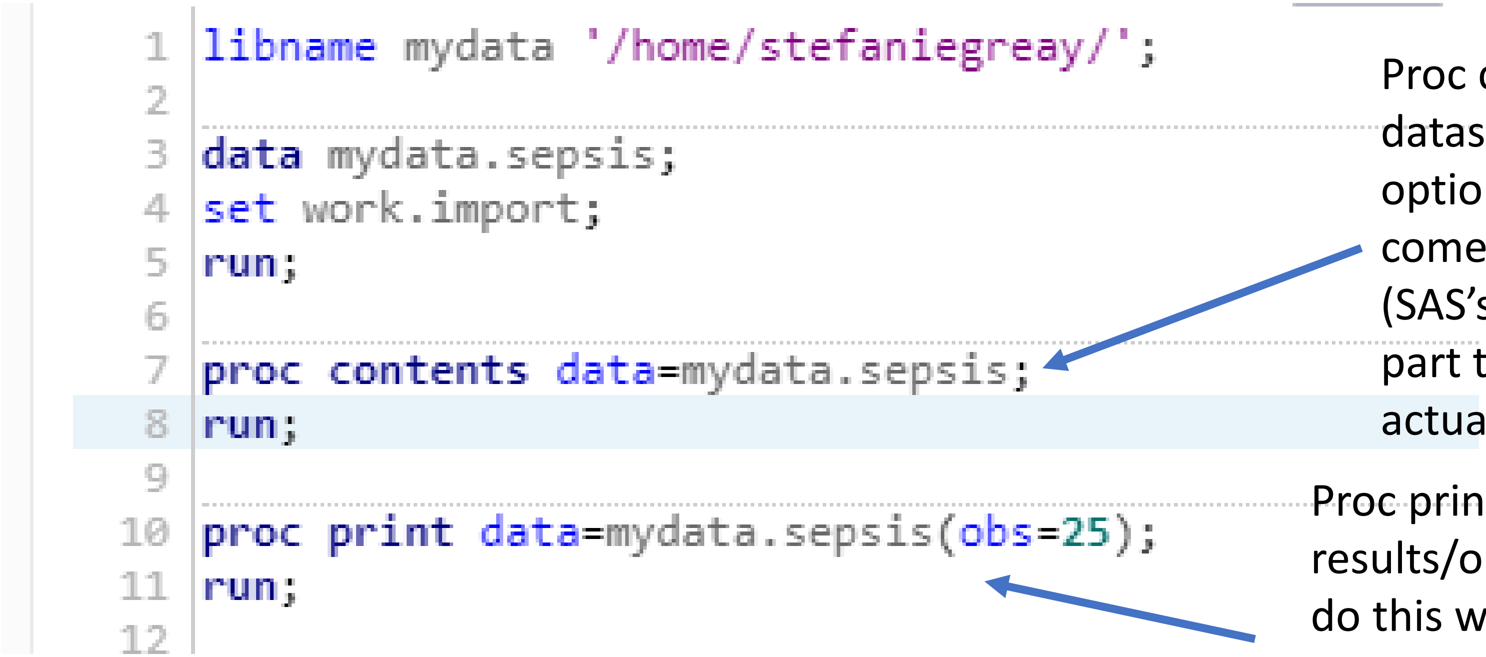
You can determine what the path is for your Files(Home) Folder by right clicking on Files(Home) and selecting Properties. (Yours will not be the same as mine – it will include your userID, not mine.)

Save the temporary SAS dataset created by the import to your library using the following sample code.



```
libname mydata '/home/stefaniegreay/';

data mydata.sepsis;
set work.import;
run;
```

First: Look at the contents of the dataset

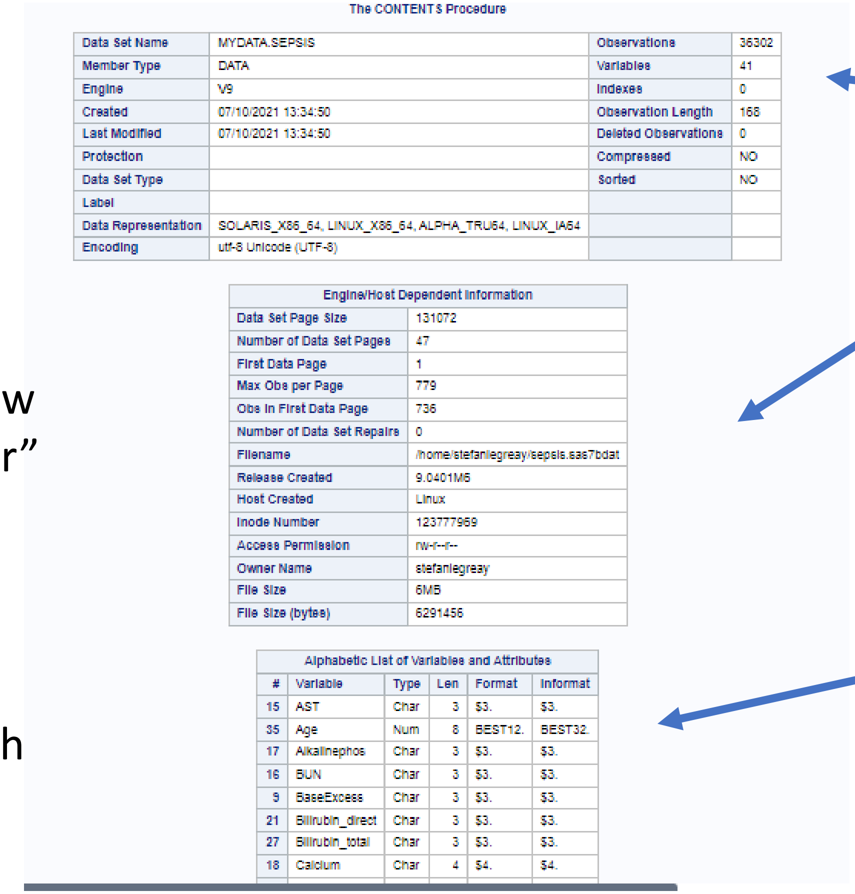


```
libname mydata '/home/stefaniegreay/';

data mydata.sepsis;
set work.import;
run;

proc contents data=mydata.sepsis;
run;

proc print data=mydata.sepsis(obs=25);
run;

```

Proc contents provides the contents of the dataset identified next to the data= option. Remember that the part that comes before the . Is the library reference (SAS's reference for the folder) and the part that comes after is the name of the actual dataset.

Proc print prints the dataset out to the results/output window. We don't want to do this without limiting the number of rows output, though, because it could crash if we have a very large dataset. We can limit the number of records/rows printed using the obs= option in parentheses. This example prints the first 25 rows.

Proc Contents Output



The CONTENTS Procedure

Data Set Name	MYDATA.SEPSIS	Observations	36302
Member Type	DATA	Variables	41
Engine	V9	Indexes	0
Created	07/10/2021 13:34:50	Observation Length	168
Last Modified	07/10/2021 13:34:50	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	47
First Data Page	1
Max Obs per Page	779
Obs in First Data Page	736
Number of Data Set Repairs	0
Filename	/home/stefaniegreay/sepsis.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	123777969
Access Permission	rw-r--r--
Owner Name	stefaniegreay
File Size	6MB
File Size (bytes)	6291456

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
15	AST	Char	3	$3.	$3.
35	Age	Num	8	BEST12.	BEST32.
17	Alkalinephos	Char	3	$3.	$3.
16	BUN	Char	3	$3.	$3.
9	BaseExcess	Char	3	$3.	$3.
21	Bilirubin_direct	Char	3	$3.	$3.
27	Bilirubin_total	Char	3	$3.	$3.
18	Calcium	Char	4	$4.	$4.

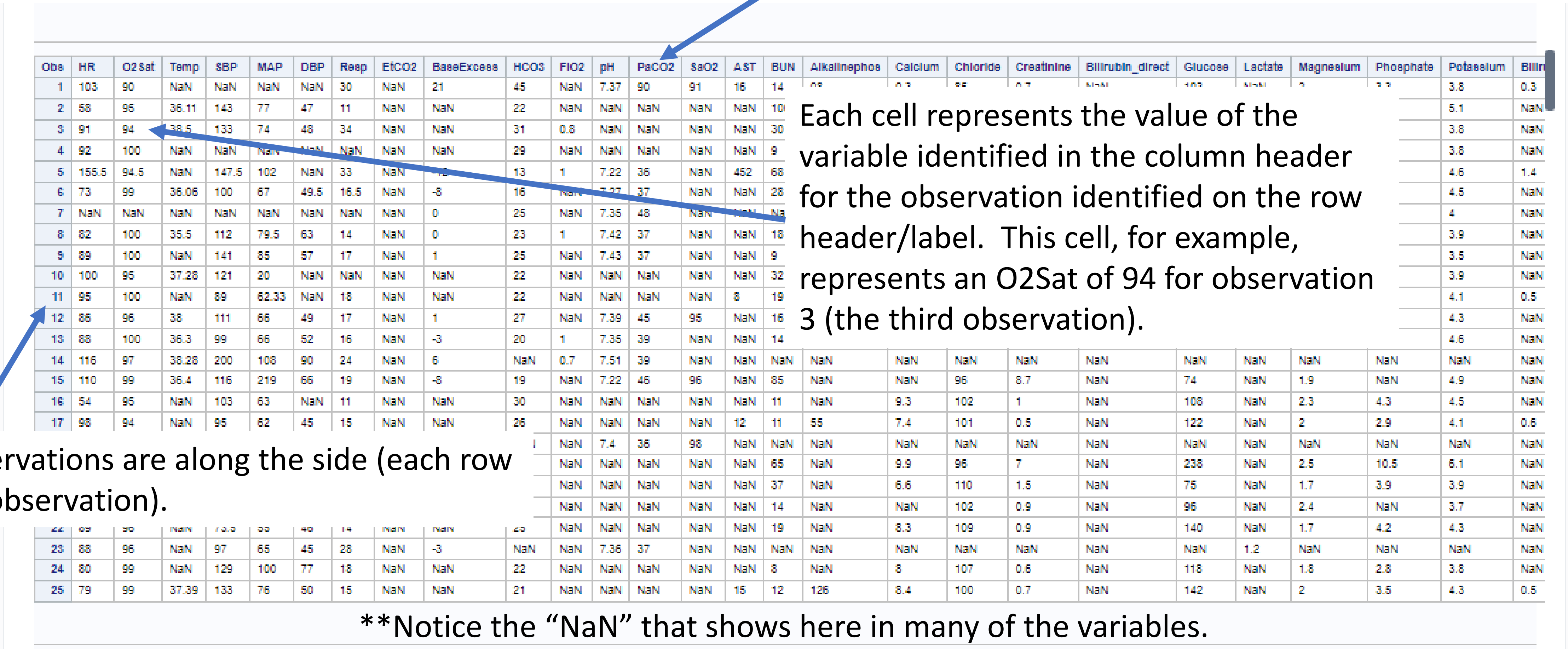
Basic dataset information is here, including the number of observations (rows) and variables (columns).

The details of all of the variables and attributes of the variables are shown here, including the variable name, variable number (order of the variable in the input file), type (only Character (Char) or Numeric (Num) variable types exist in SAS), length, format (how SAS displays the data), and informat (how the data was read in from the input file).

*Notice how many “Char” variables show up here, but with only a small length (of 3).

Proc Print Output

Variables are across the top (each column is a variable).



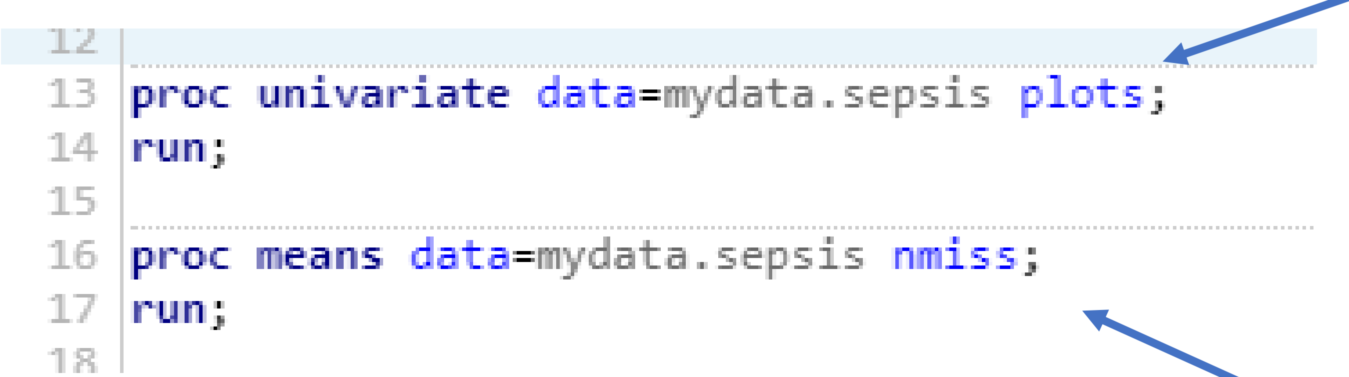
Obs	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	FiO2	pH	PaCO2	SaO2	AST	BUN	Alkalinephos	Calcium	Chloride	Creatinine	Bilirubin_direct	Glucose	Lactate	Magnesium	Phosphate	Potassium	Bilir
1	103	90	NaN	NaN	NaN	NaN	30	NaN	21	45	NaN	7.37	90	91	16	14	98	9.3	85	0.7	NaN	103	NaN	2	3.3	3.8	0.3
2	58	95	36.11	143	77	47	11	NaN	NaN	22	NaN	NaN	NaN	NaN	NaN	10										5.1	NaN
3	91	94	38.5	133	74	48	34	NaN	NaN	31	0.8	NaN	NaN	NaN	NaN	30										3.8	NaN
4	92	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	29	NaN	NaN	NaN	NaN	NaN	9										3.8	NaN
5	155.5	94.5	NaN	147.5	102	NaN	33	NaN	-12	13	1	7.22	36	NaN	452	68										4.6	1.4
6	73	99	36.06	100	67	49.5	16.5	NaN	-8	16	NaN	7.27	37	NaN	NaN	28										4.5	NaN
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0	25	NaN	7.35	48	NaN	NaN	Na										4	NaN
8	82	100	35.5	112	79.5	63	14	NaN	0	23	1	7.42	37	NaN	NaN	18										3.9	NaN
9	89	100	NaN	141	85	57	17	NaN	1	25	NaN	7.43	37	NaN	NaN	9										3.5	NaN
10	100	95	37.28	121	20	NaN	NaN	NaN	NaN	22	NaN	NaN	NaN	NaN	NaN	32										3.9	NaN
11	95	100	NaN	89	62.33	NaN	18	NaN	NaN	22	NaN	NaN	NaN	NaN	8	19										4.1	0.5
12	86	96	38	111	66	49	17	NaN	1	27	NaN	7.39	45	95	NaN	16										4.3	NaN
13	88	100	36.3	99	66	52	16	NaN	-3	20	1	7.35	39	NaN	NaN	14										4.6	NaN
14	116	97	38.28	200	108	90	24	NaN	6	NaN	0.7	7.51	39	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
15	110	99	36.4	116	219	66	19	NaN	-8	19	NaN	7.22	46	96	NaN	85	NaN	NaN	96	8.7	NaN	74	NaN	1.9	NaN	4.9	NaN
16	54	95	NaN	103	63	NaN	11	NaN	NaN	30	NaN	NaN	NaN	NaN	NaN	11	NaN	9.3	102	1	NaN	108	NaN	2.3	4.3	4.5	NaN
17	98	94	NaN	95	62	45	15	NaN	NaN	26	NaN	NaN	NaN	NaN	12	11	55	7.4	101	0.5	NaN	122	NaN	2	2.9	4.1	0.6
											NaN	7.4	36	98	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
											NaN	NaN	NaN	NaN	NaN	65	NaN	9.9	96	7	NaN	238	NaN	2.5	10.5	6.1	NaN
											NaN	NaN	NaN	NaN	NaN	37	NaN	6.6	110	1.5	NaN	75	NaN	1.7	3.9	3.9	NaN
											NaN	NaN	NaN	NaN	NaN	14	NaN	NaN	102	0.9	NaN	96	NaN	2.4	NaN	3.7	NaN
22	69	98	NaN	73.5	55	48	14	NaN	NaN	23	NaN	NaN	NaN	NaN	NaN	19	NaN	8.3	109	0.9	NaN	140	NaN	1.7	4.2	4.3	NaN
23	88	96	NaN	97	65	45	28	NaN	-3	NaN	NaN	7.36	37	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.2	NaN	NaN	NaN	NaN
24	80	99	NaN	129	100	77	18	NaN	NaN	22	NaN	NaN	NaN	NaN	NaN	8	NaN	8	107	0.6	NaN	118	NaN	1.8	2.8	3.8	NaN
25	79	99	37.39	133	76	50	15	NaN	NaN	21	NaN	NaN	NaN	NaN	15	12	126	8.4	100	0.7	NaN	142	NaN	2	3.5	4.3	0.5

Each cell represents the value of the variable identified in the column header for the observation identified on the row header/label. This cell, for example, represents an O2Sat of 94 for observation 3 (the third observation).

Observations are along the side (each row is a observation).

**Notice the "NaN" that shows here in many of the variables.

Second: Summarize the numeric variables and check for missing values



```
proc univariate data=mydata.sepsis plots;
run;

proc means data=mydata.sepsis nmiss;
run;
```

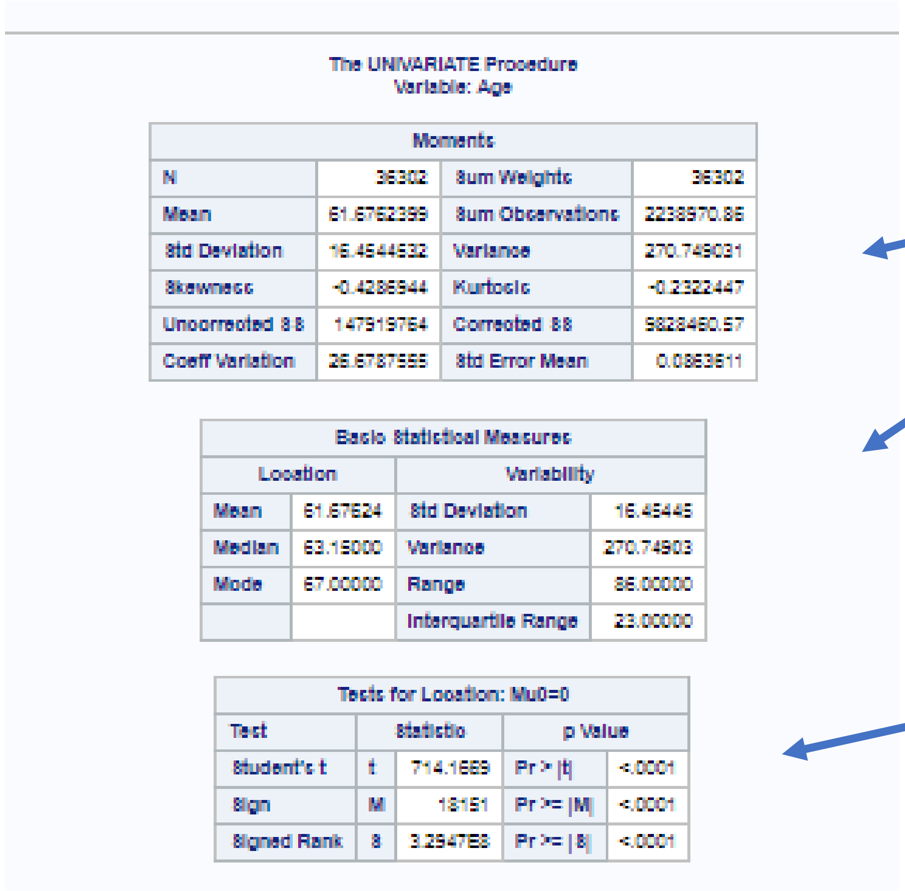
Proc univariate outputs several numeric summaries of the numeric variables, and the plots option adds several plots (the distribution plot, box plot, and normal quantile plot) for each variable. Not using a var statement tells SAS to do this summary for all variables with the type of Num (i.e. numeric variables).

Proc means with the nmiss option outputs the number of missing values for the numeric variables. Not using a var statement tells SAS to do this summary for all variables with the type of Num (i.e. numeric variables).

Sample code with var statement (to specify one or more specific numeric variables:

Proc univariate data=mydata.sepsis plots;
Var age;
Run;

Proc Univariate Output



The UNIVARIATE Procedure
Variable: Age

Moments			
N	36302	Sum Weights	36302
Mean	61.6762399	Sum Observations	2238970.86
Std Deviation	16.4544532	Variance	270.749031
Skewness	-0.4286944	Kurtosis	-0.2322447
Uncorrected SS	147919764	Corrected SS	9828460.57
Coeff Variation	26.6787555	Std Error Mean	0.0863611

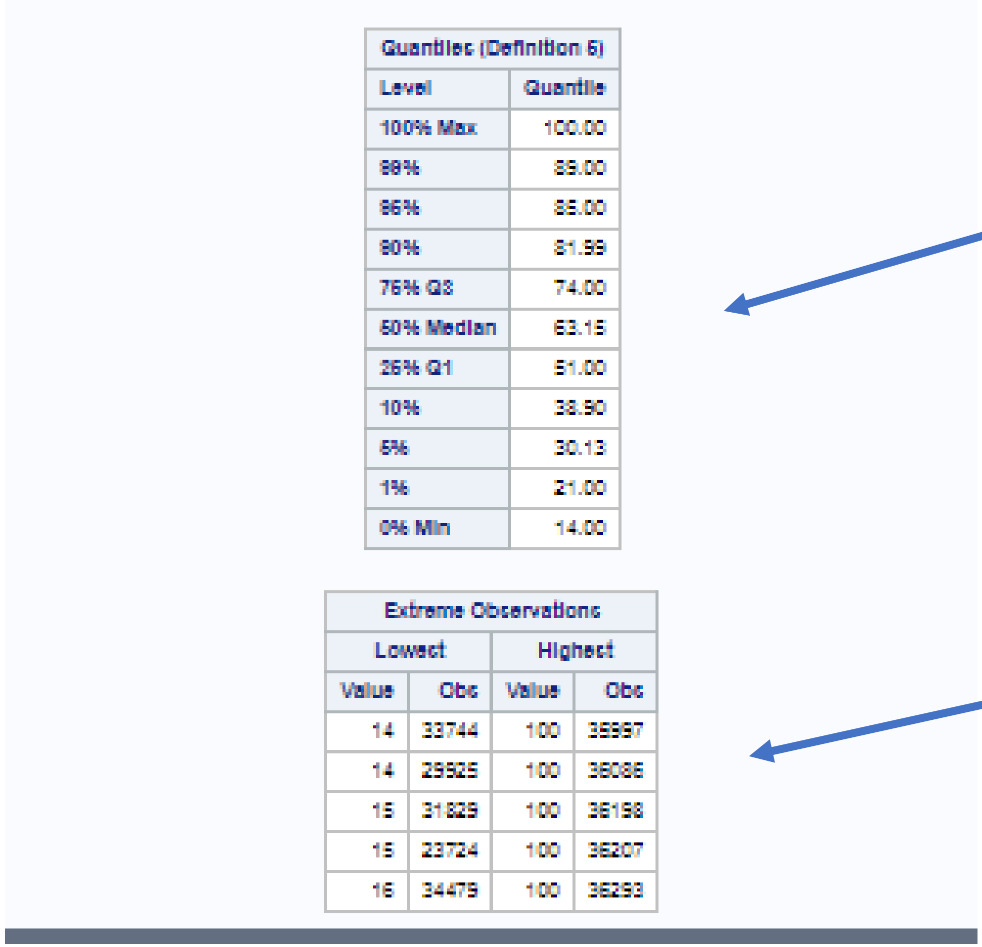
Basic Statistical Measures			
Location		Variability	
Mean	61.67624	Std Deviation	16.45445
Median	63.15000	Variance	270.74903
Mode	67.00000	Range	86.00000
		Interquartile Range	23.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	714.1669	Pr > |t|	<.0001
Sign	M	18151	Pr >= |M|	<.0001
Signed Rank	S	3.2947E8	Pr >= |S|	<.0001

Basic summary statistics (measures of central tendency/middle; measures of dispersion/spread; distribution measures (skewness; kurtosis; etc.))

This section includes hypothesis testing details for testing if the middle of the distribution is centered around 0. For most data quality checks and data profiling, this section is irrelevant.

Proc Univariate Output (Cont.)



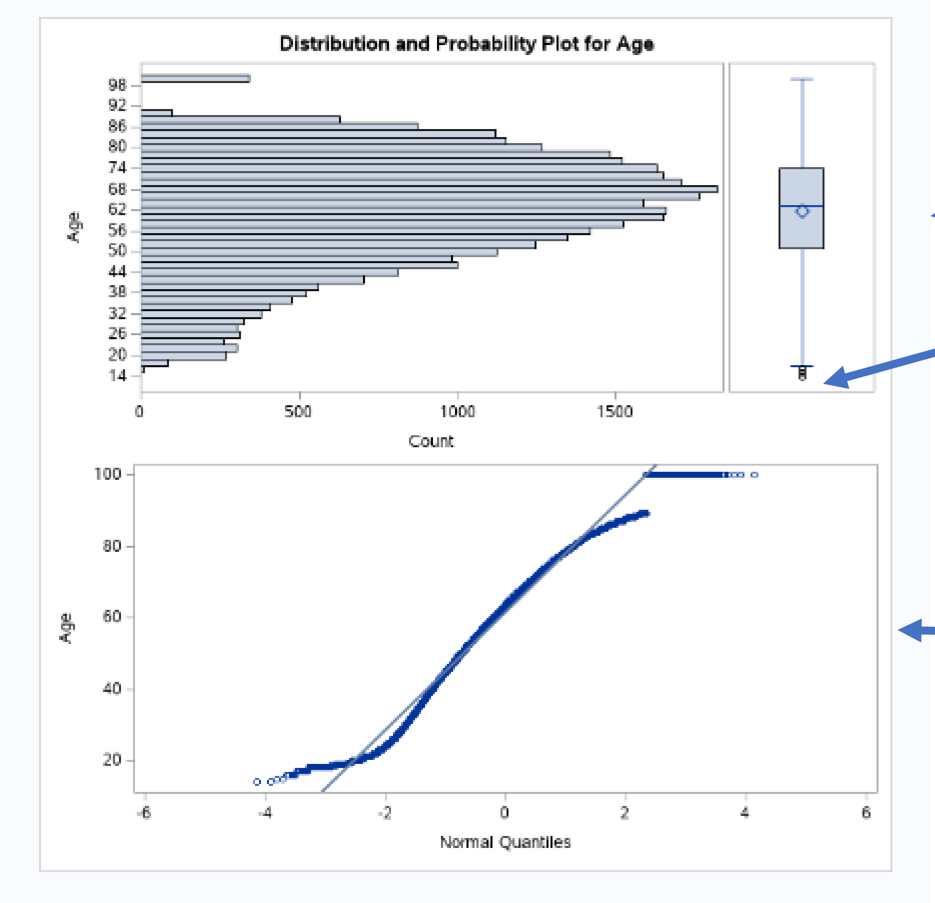
Quantiles (Definition 5)	
Level	Quantile
100% Max	100.00
99%	89.00
95%	85.00
90%	81.99
75% Q3	74.00
50% Median	63.15
25% Q1	51.00
10%	38.90
5%	30.13
1%	21.00
0% Min	14.00

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14	33744	100	35997
14	29925	100	36086
15	31829	100	36198
15	23724	100	36207
16	34479	100	36293

Percentile cutoff information for the variable is included here. For example, the 75th percentile is 74.00 in this case; the maximum is 100, and the minimum is 14.

Extreme observations are identified in this section. It includes the variable values themselves (for the lowest and highest observations), as well as the number identifying the observation (i.e. row) that that value can be found in. This is a great resource for identifying data that might have been entered incorrectly, like if someone entered 1000 for their age, instead of 100, for example.

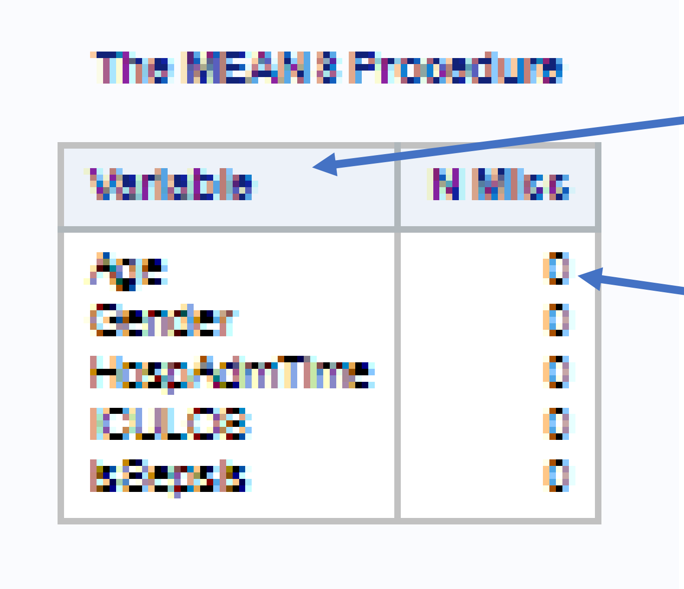
Proc Univariate Output (Cont.) - Plots



This section includes the histogram and box plot. Checking the overall shape of the distribution on these plots is a good step, as well as identifying any values that look out of place or like outliers (like the dots at the bottom of the box plot, for example).

The normal quantlies plot gives us an idea of whether or not the distribution is bell-shaped (or distributed according to a Normal distribution). The closer to all of the points following the lines, the closer to a perfect bell shape the distribution follows. Any points that stand far out from the rest may indicate ones that need to be explored further.

Proc Means (with nmiss option) Output



The MEANS Procedure

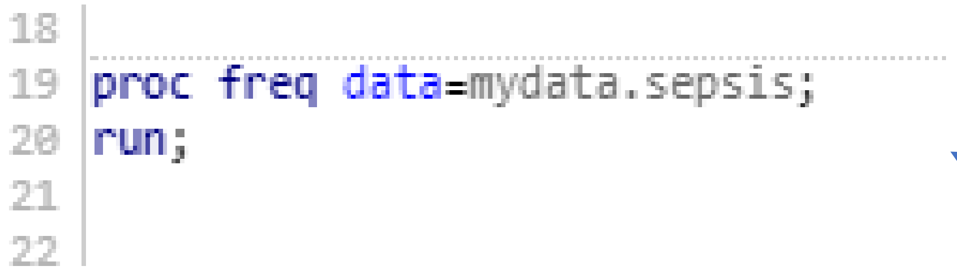
Variable	N Miss
Age	0
Gender	0
HospAdmTime	0
ICULOS	0
isSepsis	0

Variable name

Number of missing values that are included in the variable listed in the given row. (For example, there are 0 missing values shown in the Age variable for this dataset, according to this output.)

*Notice that there are only 5 variables that SAS identified as numeric and none of them have missing values, out of the 41 total variables in this dataset.

Third: Summarize the character variables and check for missing values



```
proc freq data=mydata.sepsis;
run;
```

Sample code with tables statement (to specify only using the variables formatted/recognized as character):

Proc freq data=mydata.sepsis;
Tables _CHAR_;
Run;

(To specify a specific variable, replace _CHAR_ with the name of the specific variable.)

Proc freq outputs a table that counts the frequency (number of observations) with a particular value of that variable, as well as the relative frequency (or percent of observations), cumulative frequency and cumulative relative frequency). Not using a tables statement tells SAS to do this summary for all variables. This will cause a very large set of output if we don't specify at least character values only (example to the left).

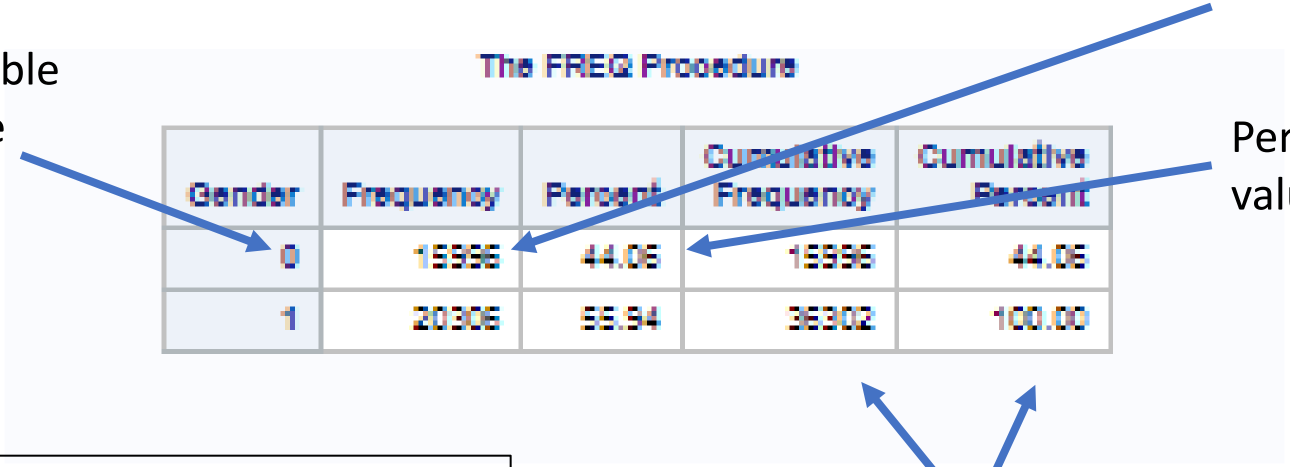
You can technically also specify a numeric variable in the tables statement, although including continuous, numeric variables (ones with a lot of different values) is not beneficial or interpretable and will cause memory issues or crash the program if the data is very large. A variable like gender, here, however, that only has values of 0 and 1 (but was imported by SAS as numeric because of the 0's and 1's would be fine to include in a proc freq.

Proc Freq Output Example

Variable value

Count of observations with that variable value (i.e. a Gender of 0 in this case)

Percent of observations with that variable value (i.e. a Gender of 0 in this case)



The FREQ Procedure

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	15996	44.06	15996	44.06
1	20306	55.94	36302	100.00

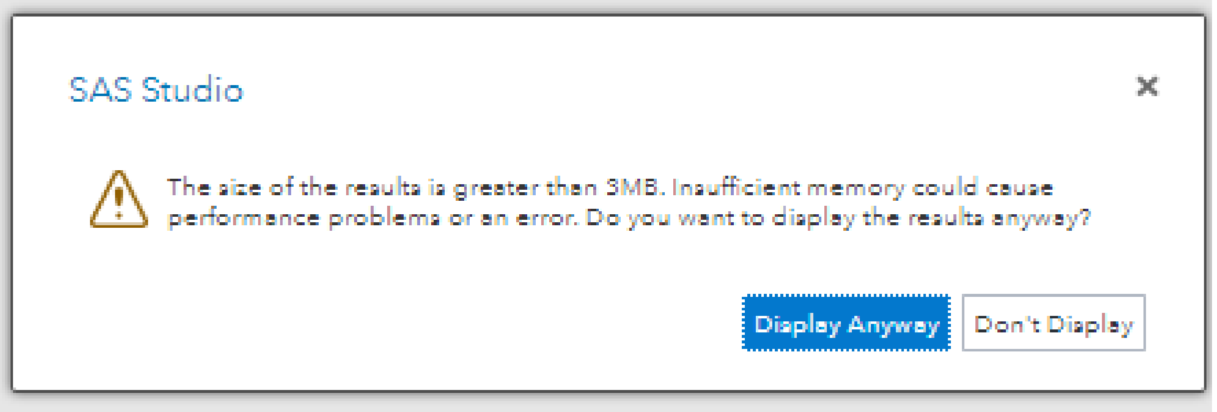
Cumulative count and percent of observations with that variable value

Code used to get this output:

```
Proc freq data=mydata.sepsis;
Tables gender;
Run;
```

A frequency table can be helpful for identifying variable values that look out of place (i.e have very small frequencies compared to the others), any spelling errors or differences in case (Fred and fred would be treated differently within SAS, as two different variable values, for example, because of the difference in capitalization, without further intervention).

Proc Freq for Variables with Many Values



SAS Studio

The size of the results is greater than 3MB. Insufficient memory could cause performance problems or an error. Do you want to display the results anyway?

This message results from the proc freq on this sepsis dataset without a tables statement. (It is trying to output one table per variable that SAS identified as character, with one row per unique variable value in each table.)

*Note that this is not a common message and indicates that something might not be quite right with the data or how it was imported or how the variable types were recognized and registered by SAS.

Code template

```
libname mydata '/home/stefaniegreay/';

data mydata.sepsis;
set work.import;
run;

proc contents data=mydata.sepsis;
run;

proc print data=mydata.sepsis(obs=25);
run;
```

```
proc univariate data=mydata.sepsis plots;
run;

proc means data=mydata.sepsis nmiss;
run;

proc freq data=mydata.sepsis;
tables gender;
run;

proc freq data=mydata.sepsis;
tables _CHAR_;
run;
```